# SR-TWAS: Leveraging Multiple Reference Panels to Improve TWAS Power by Ensemble Machine Learning

Randy L. Parrish[1,2], Aron S. Buchman[3], Shinya Tasaki[3], Yanling Wang[3], Denis Avey[3], Jishu Xu[3], Philip L. De Jager[4], David A. Bennett[3], Michael P. Epstein[1], Jingjing Yang[1*]

1. Center for Computational and Quantitative Genetics, Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, 30322, USA
2. Department of Biostatistics, Emory University School of Public Health, Atlanta, GA 30322, USA
3. Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, 60612, USA
4. Center for Translational and Computational Neuroimmunology, Department of Neurology and Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University Irving Medical Center, New York, NY10032, USA

[*] Correspondence Author: jingjing.yang@emory.edu

**Abstract** (150 words)

Multiple reference panels of a given tissue or multiple tissues often exist, and multiple regression methods could be used for training gene expression imputation models for TWAS. To leverage expression imputation models (i.e., base models) trained with multiple reference panels, regression methods, and tissues, we develop a Stacked Regression based TWAS (SR-TWAS) tool which can obtain optimal linear combinations of base models for a given validation transcriptomic dataset. Both simulation and real studies showed that SR-TWAS improved power due to increased effective training sample sizes and borrowed strength across multiple regression methods and tissues. Leveraging base models across multiple reference panels, tissues, and regression methods, our studies of Alzheimer's disease (AD) dementia and Parkinson's disease (PD) identified respective 11 independent significant risk genes for AD (supplementary motor area tissue) and 12 independent significant risk genes for PD (substantia nigra tissue), including 6 novels for AD and 6 novels for PD.

## Introduction

Two-stage transcriptome-wide association study (TWAS) has been widely used in genetics studies of complex traits, due to the convenience of using publicly available transcriptomic reference panels and summary-level genome-wide association study (GWAS) datasets[1–5]. The standard two-stage TWAS method[6,7] first trains gene expression imputation models (per gene per tissue) using a transcriptomic reference panel (Stage I), taking quantitative gene expression traits as response variables and nearby cis- or genome-wide (cis- and trans-) genetic variants as predictors. The non-zero genetic effect sizes estimated in the gene expression imputation model are considered effect sizes of a broad sense of expression quantitative trait loci (eQTL), which are taken as variant weights to conduct gene-based association tests with GWAS data (individual-level or summary-level) in Stage II.

Various TWAS techniques have been developed, employing diverse regression methods to train models for imputing gene expression. Additionally, multiple transcriptomic reference panels are made available to the public and could be used in TWAS. Consequently, it is possible to train multiple gene expression imputation models by employing distinct regression methods, employing multiple transcriptomic reference panels of the same tissue type, or utilizing transcriptomic data from multiple tissues within a given reference panel. For example, multiple regression methods, such as penalized regression with Elastic-Net penalty (used by PrediXcan[7]) and nonparametric Bayesian Dirichlet process regression (DPR) model (used by TIGAR[8]), have trained gene expression imputation models using the same Genotype-Tissue Expression (GTEx)[9] V8 reference data of 48 human tissue types. The Religious Orders Study (ROS)[10], Rush Memory and Aging Project (MAP)[10], and the GTEx[9] V8 project all profile transcriptomic data of prefrontal cortex (PFC) brain tissue and genome-wide genetic data of the same samples,

53 providing multiple reference panels of PFC tissue for TWAS. Thus, leveraging multiple trained
54 gene expression imputation models of the same target gene across multiple regression methods,
55 reference panels, and tissue types is expected to improve TWAS power, for more robustly
56 modeling the unknown genetic architecture of the target gene expression by multiple regression
57 models, having an increased training sample size with multiple reference panels, or borrowing
58 strength across multiple tissue types with correlated gene expression.

59 Multi-tissue approaches that can take advantage of transcriptomic reference data for multiple
60 tissues and/or reference panels have been developed. For example, UTMOST uses group LASSO
61 penalized multivariate regression to impute cross-tissue expression[11]. SWAM estimates a vector
62 of weights for input expression imputation models such that the weighted average of the input
63 models will give the least square error with respect to individual-level reference expression of
64 the target tissue[12]. However, these approaches have drawbacks such as being computationally
65 expensive and user-unfriendly. UTMOST requires individual-level reference data for all tissues.
66 In order to control for multicollinearity, a regularization parameter is considered by SWAM, and
67 needs to be fine-tuned based on the covariance structure of Genetically Regulated gene
68 eXpression (GReX) of all considered tissues which needs to be derived by using individual-level
69 transcriptomic data[12]. Additionally, SWAM requires that trained model input must be in the
70 same SQL database format as used for PrediXcan output[12].

71 To fill in this gap, we develop a novel TWAS method to leverage multiple summary-level gene
72 expression imputation models (i.e., base models) trained for the same target gene by the
73 ensemble machine learning technique of stacked regression[13,14]. We refer to this novel TWAS
74 method as Stacked Regression based TWAS (i.e., SR-TWAS). SR-TWAS first uses a validation
75 transcriptomic dataset of the target tissue type to optimally train a set of weights for the multiple

expression imputation base models per target gene (Stage I), by optimizing the gene expression prediction $R^2$ (i.e., the squared correlation between observed and predicted gene expression levels by the weighted average of multiple base models) in the validation dataset. Then SR-TWAS takes the weighted average eQTL effect sizes as the corresponding variant weights for gene-based association tests in Stage II. The trained expression imputation models by SR-TWAS are specific for the tissue type of the validation data, and the identified TWAS risk genes are interpreted with potential genetic effects mediated through the corresponding gene expression of the tissue type of the validation data.

With comprehensive simulation studies, we showed that expression imputation models trained by SR-TWAS had higher prediction accuracy and led to higher TWAS power than base models across all considered scenarios. SR-TWAS achieves the greatest gains in power over base models under scenarios in which a gene has a relatively higher proportion of true causal eQTL with relatively smaller eQTL effect sizes. In the real data validation and application studies using ROS/MAP and GTEx V8 reference panels and GWAS summary data of Alzheimer's disease (AD) dementia and Parkinson's disease (PD), SR-TWAS also outperformed base models trained using single reference panels and tissue types.

In the following sections, we first briefly describe the stacked regression method used by SR-TWAS. Then we describe the results of our simulation studies, validation studies using the real ROS/MAP and GTEx V8 reference panels, as well as application TWAS of AD dementia and PD. Last, we end with a discussion.

## Results

### Overview of SR-TWAS

In the framework of TWAS[7,8,15], a multivariable linear regression model is assumed for training gene expression imputation model, taking quantitative gene expression levels $E_g$ of the target gene and tissue as the response variable and cis-acting genetic variants nearby the target gene region (genotype matrix $G$) as predictors, as shown in the following formula:

$$\boldsymbol{E_g} = \boldsymbol{Gw} + \boldsymbol{\epsilon}, \qquad \epsilon_i \sim N(0,1).$$

The eQTL effect sizes $\boldsymbol{w}$ could be trained by different regression methods and/or using different reference panels.

Assume there are a total of $K$ base gene expression imputation models trained for the same target gene, with $\widehat{\boldsymbol{w}_k}, k = 1, \ldots, K$. Let $\mathbf{E}_{vg}$ denote the gene expression levels of the target gene $g$ in the target tissue type in the validation data, and $\mathbf{G}_v$ denote the genotype matrix of the same genetic predictors in the validation data. Then the predicted GReX of the validation samples by the $k$th base model are given by $\mathbf{G}_v\widehat{\mathbf{w}}_k$. The stacked regression method[13,14] will solve for a set of optimal base model weights $\zeta_1, \ldots, \zeta_K$, by maximizing the regression $R^2$ between the profiled gene expression $\mathbf{E}_{vg}$ and the weighted average GReX, $\sum_{k=1}^{K} \zeta_k \, \mathbf{G}_v\widehat{\mathbf{w}}_k$, of $K$ base models, i.e., minimizing the following loss function of $1 - R^2$:

$$\text{minimize}_{(\zeta_k;\, k=1,\ldots,K)} \frac{\left\|\mathbf{E}_{vg} - \sum_{k=1}^{K} \zeta_k \, \mathbf{G}_v\widehat{\mathbf{w}}_k\right\|^2}{\left\|\mathbf{E}_{vg} - E_{vg}\right\|^2}, \quad \text{s.t.} \sum_{k=1}^{K} \zeta_k = 1, \; \zeta_k \in [0,1].$$

As a result, we will obtain a set of model weights $\zeta_k$ for $k = 1, \ldots, K$ base models, and a set of eQTL effect sizes $\widetilde{\mathbf{w}}$ given by the weighted average of the eQTL effect sizes of $K$ base models, $\widetilde{\mathbf{w}} = \sum_{k=1}^{K} \zeta_k \, \widehat{\mathbf{w}}_k$ (Stage I). Then the final predicted GReX for test genotype data $\mathbf{G}_t$ is given by $\widehat{\text{GReX}}_g = \mathbf{G}_t\widetilde{\mathbf{w}}$, and $\widetilde{\mathbf{w}}$ will be taken as variant weights in the gene-based association tests by

SR-TWAS in Stage II. Genes with 5-fold cross validation (CV) $R^2 > 0.5\%$ in the validation dataset by SR-TWAS are considered as having a valid imputation model and will be tested in Stage II[8]. Here, $\tilde{\mathbf{w}}$ is the trained eQTL effect sizes by SR-TWAS (Stage I) for the target gene of the tissue of the validation data, and identified significant genes from Stage II have potential genetic effects mediated through the transcriptome of the tissue of the validation data.

**Simulation Studies**

By simulation studies, we compared the performance of SR-TWAS with a Naïve approach which takes the average of base models as the trained gene expression imputation model, that is, taking $\zeta_k = \frac{1}{K}, k = 1, \ldots, K$. We used the real genotype data of gene *ABCA7* from ROS/MAP and GTEx V8 to simulate gene expression and phenotypes, and considered multiple scenarios with varying proportions of causal SNPs ($p_{causal} = (0.001, 0.01, 0.05, 0.1)$) and gene expression heritability (i.e., the proportion of gene expression variation due to genetics, $h_e^2 = (0.1, 0.2, 0.5)$). We randomly selected n=465 training samples with Whole Genome Sequencing (WGS) genotype data from the ROS/MAP cohort and GTEx V8 cohort, respectively. We randomly selected n=400 and n=800 samples with WGS genotype data from ROS/MAP as our validation and test cohorts, respectively. Training, validation, and test samples from the ROS/MAP cohort were simulated with the same causal SNPs (i.e., eQTL), while training samples from the GTEx V8 cohort were simulated with true causal SNPs that were 50% overlapped with the ones for ROS/MAP samples. The simulated expression heritability was the same for both ROS/MAP and GTEx V8 samples.

Two base models per gene were trained by PrediXcan (penalized regression with Elastic-Net penalty) with the GTEx training samples (n=465), and by TIGAR (nonparametric Bayesian Dirichlet process regression) with the ROS/MAP training samples (n=465). SR-TWAS and Naïve models were then obtained by using these trained base models. Validation data ($n = 400$) were used to train SR-TWAS models. Gene expression imputation models (by SR-TWAS and Naïve methods) with 5-fold cross-validation $R^2 > 0.5\%$ in the validation cohort were considered valid models and used for follow-up gene-based association tests. Test data (n=800) were used for assessing GReX prediction performance and TWAS power, with 1,000 repeated simulations per scenario. We compared the performance of SR-TWAS, Naïve method, and these two base models.

As shown in **Fig. 1**, we showed that SR-TWAS obtained the highest test $R^2$ for gene expression imputation across all scenarios, with slightly better performance than the base models trained by TIGAR with ROS/MAP samples (TIGAR_ROSMAP), by leveraging the predictive information provided by both base models. Both SR-TWAS and the TIGAR_ROSMAP base models performed better than the Naïve method. The base models trained by PrediXcan with GTEx samples (PrediXcan_GTEx) performed the worst. This is because all ROS/MAP training, validation, and test samples are simulated under the same causality model with the same set of causal SNPs, while GTEx training samples were simulated only with 50% overlapped true causal SNPs of the validation and test samples. The Naïve approach of taking averages of the base models had poor performance because of the heterogeneous genetic architecture between the GTEx training cohort and test cohort. As expected, model performance improved with increasing true expression heritability $h_e^2$ with the same training sample size. For all considered scenarios, the highest test $R^2$ were obtained under a sparse causality model with $p_{\text{causal}} = 0.001$, where true

causal SNP effect sizes would be relatively larger given the same $h_e^2$. CV $R^2$ and training $R^2$ results for Naïve and SR-TWAS approaches for these scenarios (Supplementary Figures 1-2) also showed that SR-TWAS outperformed the Naïve approach under all scenarios.

In order to assess TWAS power, phenotypes were simulated with a certain proportion of variance due to simulated gene expression ($h_p^2$). We considered a series of $h_p^2$ values in the range of $(0.05, 0.875)$. The TWAS power comparison by SR-TWAS, naïve method, and two base models were shown in **Fig. 2**, where the results were consistent with the test $R^2$ comparison as in **Fig 1**. SR-TWAS performed similarly to the TIGAR_ROSMAP base model, while the Naïve method performed worse than SR-TWAS and TIGAR_ROSMAP base model but still better than the PrediXcan_GTEx base model. In particular, SR-TWAS had a noticeable advantage over the TIGAR_ROSMAP model with $p_{\text{causal}} = 0.01$ and $h_e^2 = 0.5$. Although desirable TWAS power ~80% was only obtained in simulation scenarios with a relatively high $h_p^2$ that might be higher than the value in real studies, simulation power would increase along with increased test sample sizes. Because real GWAS test data would have a larger sample size than the 800 considered in our simulations, we expect desirable power for our SR-TWAS method in real studies.

Additionally, we conducted similar simulation studies for two other settings, where samples from ROS/MAP and GTEx cohorts have the same set of true causal SNPs (i.e., the same genetic architecture), and (i) the expression heritability was the same for both ROS/MAP and GTEx V8 cohorts, or (ii) the expression heritability for GTEx V8 cohort is only half that of ROS/MAP. The results of these settings were similar to that of the previously described setting. SR-TWAS and TIGAR_ROSMAP models outperformed the PrediXcan-GTEx and Naïve methods. Comparisons of CV $R^2$ and training $R^2$ for Naïve and SR-TWAS approaches for these scenarios (Supplementary Figures 3, 4, 7, and 8) showed that SR-TWAS outperformed the Naïve approach

under all scenarios. For all considered scenarios, again the highest test $R^2$ was obtained under a sparse causality model with high expression heritability (Supplementary Figures 5 and 9). Power comparison results show that SR-TWAS and TIGAR_ROSMAP models generally outperformed the PrediXcan-GTEx and Naïve methods (Supplementary Figures 6 and 10), particularly in the setting in which the expression heritability for GTEx V8 cohort was only half that of ROS/MAP (Supplementary Figure 10).

We also assessed type I error under the example scenario with $p_{\text{causal}} = 0.1, h_e^2 = 0.1$. Base model weights were permuted $10^6$ times and used to train SR-TWAS and Naïve models, which were then used to conduct gene-based association tests with a phenotype generated randomly from $N(0,1)$. All methods control well for type I errors for significance thresholds $(10^{-4}, 10^{-5}, 2.5 \times 10^{-6}, 10^{-6})$, as shown in Supplementary Table 1. Type I errors for Naïve method were comparable to that of the base models, while SR-TWAS had the lowest type I error under all significance thresholds.

**Real Validation Studies**

To compare the GReX prediction accuracy with real gene expression data, we considered three base models that were trained by TIGAR with ROS samples (n=237, TIGAR_ROS_DLPFC) for dorsolateral prefrontal cortex (DLPFC) tissue, trained by TIGAR with GTEx V8 data of the brain frontal cortex tissue[8] (n=157, TIGAR_GTEx_BRNCTXB), and trained by PrediXcan with the same GTEx reference data of brain frontal cortex tissue (n=157, PrediXcan_GTEx_BRNCTXB)[7]. SR-TWAS (SR-TWAS_MAP_DLPFC) and Naïve (Naive_MAP_DLPFC) models were trained from these three base models with respect to a

validation dataset with half MAP samples (n=114, randomly selected) of DLPFC tissue. Valid gene expression imputation models trained by SR-TWAS and Naïve methods with 5-fold CV $R^2 > 0.5\%$ in validation data were tested using the other half MAP samples (n=114) of DLPFC tissue.

By comparing test $R^2$ obtained by SR-TWAS, Naïve, and three base models (Supplementary Table 2), we showed that PrediXcan_GTEx_BRNCTXB had the highest median (0.070) and mean (0.113) test $R^2$ but only for 867 valid gene expression imputation models, SR-TWAS had the second highest median (0.026) and mean (0.068) test $R^2$ for 8425 valid genes expression imputation models, and Naïve model performed similarly as SR-TWAS but with a slightly lower median (0.025) and mean (0.065) test $R^2$ and fewer valid genes expression imputation models (8360). By pair-wise comparison of test $R^2$ for all genes with valid expression imputation models as shown in Supplementary Figure 11, SR-TWAS (y-axis) performed noticeably better than Naïve and three base models (x-axis).

## Application TWAS of AD Dementia

### Training expression imputation models of SMA tissue by SR-TWAS

We considered four base models — TIGAR and PrediXcan models trained with 465 ROS/MAP samples of DLPFC tissue (TIGAR_ROSMAP_DLPFC, PrediXcan_ROSMAP_DLPFC), TIGAR and PrediXcan models trained with 157 GTEx V8 samples of prefrontal cortex tissue (TIGAR_GTEx_BRNCTXB, PrediXcan_GTEx_BRNCTXB). Additional 76 ROS/MAP samples of the supplementary motor area (SMA) brain tissue were used as the validation dataset to train SR-TWAS models and to calculate the 5-fold CV $R^2$ that was used to select genes with valid

imputation models. Here, we compared SR-TWAS to base models trained by two different regression methods as used by TIGAR and PrediXcan, as well as models trained by TIGAR using the validation data of the target SMA brain tissue (TIGAR_ROSMAP_SMA), to show the advantages of SR-TWAS about leveraging multiple regression models, reference panels, and multiple tissues.

By comparing the CV $R^2$ and numbers of genes with valid expression imputation models obtained by SR-TWAS and 4 base models (**Table 1**), we found that gene expression imputation models trained by SR-TWAS for the SMA tissue (SR-TWAS_ROSMAP_SMA) had the highest median CV $R^2$ (~0.07) and second highest mean CV $R^2$ (0.09) for ~20K genes with valid expression imputation models. Although the PrediXcan_GTEx_BRNCTXB base model had third highest median CV $R^2$ (0.061) and highest mean CV $R^2$ (0.10), but only for 4563 genes with valid expression imputation models. The TIGAR_ROSMAP_SMA models with the lowest sample size of the single-cohort models (n=76), had the greatest number of trained genes (32350), second highest median CV $R^2$ (0.064), and second highest mean CV $R^2$ (0.076), which is consistent with overfitting trends in model training as observed in previous studies[8]. These results showed that improved CV $R^2$ in a real validation cohort of SMA tissue was obtained by SR-TWAS, by leveraging multiple regression methods from two reference panels of multiple relevant tissues (DLPFC and frontal cortex tissues).

**TWAS results of AD dementia**

By using the eQTL weights obtained by SR-TWAS using the above four base models and SMA validation data and TIGAR models trained using the SMA validation data, we conducted TWAS with the summary-level data of the most recent GWAS of AD dementia (n=~762K)[16]. SR-

TWAS identified a total of 56 significant TWAS risk genes of AD dementia with p-values < $2.5 \times 10^{-6}$. Of these, 17 are known GWAS risk genes, 39 are within 1MB of a known GWAS risk gene, and 19 have been previously identified as TWAS risk genes of AD dementia[17–22] (Supplementary Table 3).

Because TWAS considers genotype data within a $\pm$1MB region of a test gene, nearby significant TWAS genes with overlapping test regions often have correlated GReX values and might not represent independent associations. We curated 11 independent TWAS risk genes of AD from these 56 significant genes (**Table 2, Fig. 3**), by selecting the most significant gene as the independent risk gene for a cluster of significant genes with overlapped test regions. We found that 6 of these independent risk genes were novel TWAS risk genes (*AC073842.1, DMPK, FAM13C, GFAP, PPP1R9B* , and *SLC15A3*), 4 were known GWAS risk genes (*ACE*[18,23,24], *CR1*[16,18,19,24,25], *HLA-DRA*[19,24], and *TREM2*[16,19,24,26]), and 5 were previously identified as TWAS risk genes (*ACE*[17], *CR1*[17], *HLA-DRA*[18], *TREM2*[21], and *ZSCAN26*[18]). Importantly, 3 out of these 11 independent risk genes were also identified by TIGAR with ROSMAP SMA validation data (TIGAR_ROSMAP_SMA). Compared to the TWAS results using individual base models (Supplementary Figure 12), both SR-TWAS and TIGAR_ROSMAP_SMA models identified a greater number of independent risk genes.

Interestingly, *ACE* is a protein-coding gene involved in the regulation of blood pressure and cerebral blood flow. Gene *CR1* encodes a complement system protein that may play a role in amyloid beta (an important AD pathology) clearance[25], which was also identified by previous Mendelian randomization studies[18]. Gene *HLA-DRA* is located in the major histocompatibility complex region that is expressed in glial cells[27] and which has also been previously identified by eQTL analysis[19]. Gene *TREM2*, which encodes membrane receptor in microglia and other

immune cells and may be related to chronic inflammation[26], is also a known risk gene of AD identified by eQTL colocalization[24]. Gene *ZSCAN26*, near the known GWAS risk gene *OR2B2*[28] and identified by previous TWAS[18], is a protein-coding gene predicted to be involved in DNA-binding of transcription factors and regulation of transcription.

The other 6 independent significant TWAS risk genes (*AC073842.1*, *DMPK*, *FAM13C*, *GFAP*, *PPP1R9B*, and *SLC15A*) identified by SR-TWAS are within 1MB of known GWAS risk genes[16,18,19,24,28,29], with the former 5 also being within 1MB of a previously identified TWAS risk gene[17,18] and the latter identified by TWAS[18]. *AC073842.1* is a long non-coding RNA gene located near GWAS risk gene *AP4M1*[18] (also identified by SR-TWAS as a significant TWAS risk gene), and known TWAS risk gene *PMS2P1*[17]. SR-TWAS also identified 8 other genes in this region (see Supplemental Table 3), all of which were also within 1MB of *PMS2P1*. *DMPK* was the most significant association of a cluster of 29 genes in the same overlapped test region identified by SR-TWAS, whose test region overlaps with the well-known GWAS risk gene *APOE*[16,18,19,23,25,28]. A pathogenic repeat expansion in a noncoding region of *DMPK* causes myotonic dystrophy 1 (DM1), a multi-system disease which includes brain involvement[30]. Studies of brain pathology in DM1 patients have found associations between *DMPK* and the AD-related genes *MAPT*, *APP*, and *SNCA*[30]. *FAM13C* is a protein-coding gene which has been found to be an independent prognostic marker in prostate cancer[31], a malignancy associated with AD[32]. The protein product of *GFAP* is expressed on glial cells and has been used as a biomarker of astrocyte activation and inflammation[27,33]. Multiple studies of *GFAP* note its potential as a prognostic biomarker for predicting future dementia[33]. Pereira, et al found that plasma levels of GFAP protein had greater performance as a predictive biomarker of amyloid-β positivity than the TREM2 protein[33]. *PPP1R9B* encodes the protein spinophilin that is expressed in dendritic spines

and linked to AD progression[34,35]. Further, mouse models of AD have found *TREM2* to be an important regulator of spinophilin expression[36].

Additionally, *SLC15A3* is located in the AD-associated *MS4A* gene cluster[29], which contains multiple known GWAS risk genes[16,18–20] as well as TWAS risk gene *MS4A2*[17] of AD. The *MS4A* gene cluster is notable due its role in the regulation of soluble *TREM2* in cerebrospinal fluid in AD[29]. *SLC15A3* encodes a protein that enables transmembrane transport of histidine and di-/tripeptides across cell membranes and may play a role in the innate immune response and inflammation[37]. It has been recently identified as a differentially expressed gene in AD[22].

## Application TWAS of PD

### Training expression imputation models of brain substantia nigra tissue by SR-TWAS

We considered six base models trained by TIGAR on six different tissues from GTEx V8 — brain anterior cingulate cortex BA24 (BRNACC) (n=136), brain caudate basal ganglia (BRNCDT) (n=173), brain cortex (BRNCTXA) (n=184), brain nucleus accumbens basal ganglia (BRNNCC) (n=182), brain putamen basal ganglia (BRNPTM) (n=154), and whole blood (BLOOD) (n=574). An additional 101 GTEx samples of brain substantia nigra (BRNSNG) tissue were used as the validation data to train SR-TWAS models and to calculate the 5-fold CV $R^2$ that was used to select genes with valid expression imputation models. TIGAR models were also trained on the validation data of brain substantia nigra tissue (TIGAR_GTEx BRNSNG) to compare TWAS results with that of SR-TWAS.

### TWAS Results of PD

Besides AD, we also conducted TWAS using GWAS summary statistics by the most recent biggest studies of PD (n=~33K cases, ~18k UK Biobank proxi-cases, and ~828K controls)[38], where the eQTL weights obtained by the above SR-TWAS models based on six base models of multiple tissues and validation data of brain substantia nigra tissue and TIGAR_GTEx_BRNSNG models were used. As a result, SR-TWAS identified a total of 62 significant TWAS risk genes of PD. Of these, 15 are known GWAS risk genes, 42 are within 1MB of a known GWAS risk gene, and 17 have been previously identified as TWAS risk genes of PD (Supplementary Table 4).

Similarly, from these 62 risk genes, we curated 12 independent TWAS risk genes of PD (**Fig. 4; Table 3**), including 5 novel TWAS risk genes (*IDUA*, *LA16c-385E7.1*, *LRRC37A4P*, *SHROOM3*, and *SLC30A3*). Of these novel TWAS risk genes, three (*IDUA*, *LRRC37A4P*, *SHROOM3*) are near known GWAS risk genes (*GAK*[38], *FAM47E*[38], *MAPT*[38]). The other 7 known TWAS risk genes were also known GWAS risk genes (*CD38*[39,40], *GPNMB*[38,40], *MMRN1*[38,39], *NDUFAF2*[38], *RAB29*[38,40], *VKORC1*, and *ZSWIM7*[39]). Importantly, 2 of these independent significant TWAS risk genes (*GPNMB*[38,40], *MMRN1*[38,39]) were also identified by the TIGAR_GTEx_BRNSNG models. Compared to the TWAS results using these six base models (Supplementary Figure 13), SR-TWAS models still identified the greatest number of independent risk genes while the TIGAR_GTEx_BRNSNG models identified the fewest.

Defects in *IDUA* are known to cause the lysosomal storage disorder Hurler syndrome; lysosomal mechanisms are thought to play a role in PD pathogenesis; and lysosomal storage disorder gene variants have been associated with increased PD risk[41]. *NDUFAF2* encodes for a component of mitochondrial complex I and loss of its functionality results in a rare mitochondrial encephalopathy with frequent substantia nigra pathology and motor symptoms[42]. *NDUFAF2* was

also identified as a potential drug target in a Mendelian randomization study of potential drug targets for PD treatment[43]. *SHROOM3* is involved in neural tube development, cell shape, and epithelial morphogenesis and variants which alter its expression have been implicated in chronic kidney disease[44], which shares pathophysiological mechanisms with PD and has been shown to increase PD risk[45]. Synaptic Zn2+ has been implicated in the pathophysiology of PD and the only known transporter of Zn2+ into synaptic vesicles, ZnT3, is the protein product of *SLC30A3*[46]. *ZSWIM7* has been identified as a TWAS risk gene[39]. The protein product of *VKORC1* is an important component of normal blood coagulation and is targeted by the anticoagulant drug warfarin[43]. *VKORC1* was also identified by Mendelian randomization as a potential drug target in PD treatment[43].

*LRRC37A4P* is a pseudogene near a known TWAS risk gene *LRRC37A2*[39] (also identified by SR-TWAS). *CD38* is involved in neurodegeneration, neuroinflammation, and aging[39,47]. *GPNMB* codes for a glycoprotein observed upon tissue damage and inflammation[48], and the *GPNMB* protein has been found to be elevated in PD patients after lysosomal stress[48]. *MMRN1* is a carrier protein for platelet factor V and lies ~84KB downstream of a well-established GWAS risk locus found in multiple populations[38]. *RAB29* has been implicated as a regulator of PD-associated *LRRK2*[49].

## Discussion

We present a novel TWAS tool (SR-TWAS) using the ensemble machine learning technique of stacked regression [13,14,50], for leveraging multiple gene expression imputation models trained by different regression methods and/or using different transcriptomic reference panels of different

tissue types. We demonstrated the advantages of SR-TWAS through comprehensive simulation studies, testing gene expression prediction accuracy in real data, and real TWAS of AD dementia and PD. We showed that SR-TWAS outperformed the Naïve method, base models, and models trained using a single reference panel of the target tissue, especially when base models are trained by diverse regression methods and reference panels of different tissue types. Our results show similar strength to previous studies using stacked regression in other fields[14].

Especially, in the real application TWAS of AD dementia and PD, SR-TWAS identified a greater number of total independent risk genes than any of the base models and similar or greater number than models trained using a single reference panel of the target tissue. Besides known GWAS/TWAS risk genes, or nearby known GWAS/TWAS risk genes, that were identified by SR-TWAS, we also found 6 novel independent TWAS risk genes for AD dementia and 6 novel independent TWAS risk genes for PD with known functions in respective disease pathology. Additionally, we found interesting biological interpretations relevant to AD dementia and PD for our identified TWAS risk genes. These application TWAS results indeed showed the robust and practical useful performance of our SR-TWAS tool.

The SR-TWAS tool, including the Naïve method as an option, is publicly available on GitHub. The SR-TWAS tool implements user-friendly features, including accepting genotype data of standard VCF-format as input, enabling parallel computation, and using efficient computation strategies to reduce time and memory usage. The most computation expensive part is to train all base models with different reference panels, which is subject to the regression method. For example, with training sample size n=465, PrediXcan (Elastic-Net) costs ~1 CPU minute and TIGAR (DPR) costs ~3 CPU minutes in average per gene. Publicly available trained models can also be used as base models by the SR-TWAS tool. The process of training SR-TWAS models

from base models and validation data is quite computational efficient. For example, with the ROS/MAP SMA tissue validation dataset (n=76) and four base models in our real studies, SR-TWAS model training costs ~15 CPU seconds per gene. With the GTEx substantia nigra tissue validation dataset (n=101) and six base models in our real studies, SR-TWAS model training costs ~103 CPU seconds per gene.

SR-TWAS still has its limitations. For example, SR-TWAS only considers cis-eQTL during model training, uses the standard two-stage TWAS, requires an additional validation dataset of the target tissue independent of those used for base model training[14], and assumes samples of the validation dataset and test GWAS cohort are of the same ancestry[51]. Previous studies have illustrated the importance of considering both cis- and trans- eQTL in TWAS[52], and a joint modeling of the gene expression imputation and the gene-based association test[53 54]. The stacked regression technique used by SR-TWAS also applies to scenarios considering both cis- and trans- eQTL, when base models trained with both cis- and trans- eQTL are available. Using the recently developed variance component gene-based association test[55] in Stage II is expected to help account for the uncertainty of eQTL effect sizes estimation in the reference panel, a common issue for the two-stage TWAS methods.

Overall, the SR-TWAS tool provides a useful resource for researchers to take advantages of the publicly available gene expression imputation models by using multiple regression methods (e.g., PrediXcan[7], FUSION[15], TIGAR[8]) and different reference panels of multiple tissue types (e.g., ROS/MAP[10], GTEx[9]). In particular, the final trained gene expression imputation model by SR-TWAS will be with respect to the same tissue type as the validation data set. Because multiple base models would not only increase the robustness of the gene expression imputation

406 model but also increase the total effective training sample size, SR-TWAS is expected to further
407 increase TWAS power for studying complex human diseases.

408

409 **Acknowledgements**

410 RP and JY are supported by National Institutes of Health (NIH/NIGMS) grant award
411 R35GM138313. MPE was supported by NIH/NIGMS grant award R01GM117946 and NIH/NIA
412 grant award RF1AG071170. ROS/MAP study data were provided by the Rush Alzheimer's
413 Disease Center, Rush University Medical Center, Chicago, IL. Data collection was supported
414 through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146,
415 R01AG36836, R01AG56352, U01AG32984, U01AG46152, U01AG61356, the Illinois
416 Department of Public Health, and the Translational Genomics Research Institute.

**Table 1. Comparison of CV $R^2$ of SMA tissue for valid gene expression imputation models trained by SR-TWAS, TIGAR with validation data, and four base models with ROS/MAP and GTEx V8 reference panels of DLPFC tissue.**

| | Sample Size | Median CV $R^2$ | Mean CV $R^2$ | $N_{genes}$ |
|---|---|---|---|---|
| PrediXcan_GTEx_BRNCTXB[a] | 157 | 0.061 | 0.100 | 4563 |
| PrediXcan_ROSMAP_DLPFC[a] | 465 | - | - | 6532 |
| TIGAR_GTEx_BRNCTXB[a] | 157 | 0.038 | 0.065 | 21921 |
| TIGAR_ROSMAP_DLPFC[a] | 465 | 0.016 | 0.048 | 11981 |
| TIGAR_ROSMAP_SMA | 76 | 0.064 | 0.076 | 32350 |
| SR-TWAS_SMA | - | 0.072 | 0.090 | 20216 |

*a* base model used by SR-TWAS

**Table 2. Independent TWAS risk genes of AD dementia identified by SR-TWAS.**

| | | | | SR-TWAS SMA | | TIGAR SMA | |
|---|---|---|---|---|---|---|---|
| **Gene** | **CHR** | **Start** | **End** | **Zscore** | **Pvalue** | **Zscore** | **Pvalue** |
| CR1[ac] | 1 | 207496147 | 207641765 | -7.30 | 2.91e-13 | -2.64 | 8.17e-03 |
| ZSCAN26[bc] | 6 | 28267010 | 28278224 | 4.71 | 2.50e-06 | -3.56 | 3.69e-04 |
| HLA-DRA[ac] | 6 | 32439878 | 32445046 | -6.88 | 5.93e-12 | -0.00 | 9.98e-01 |
| TREM2[ac] | 6 | 41158506 | 41163186 | 5.42 | 6.05e-08 | 4.00 | 6.38e-05 |
| AC073842.1[bd] | 7 | 100130964 | 100140439 | 5.86 | 4.54e-09 | -5.09 | 3.65e-07 |
| FAM13C[bd] | 10 | 59246130 | 59363181 | -4.87 | 1.13e-06 | 3.49 | 4.83e-04 |
| SLC15A3[bd] | 11 | 60937084 | 60952530 | -7.57 | 3.82e-14 | -4.74 | 2.11e-06 |
| GFAP[bd] | 17 | 44903159 | 44916937 | -4.73 | 2.21e-06 | -1.03 | 3.05e-01 |
| PPP1R9B[bd] | 17 | 50133737 | 50150677 | 5.11 | 3.15e-07 | -3.59 | 3.25e-04 |
| ACE[ac] | 17 | 63477061 | 63498380 | 5.05 | 4.44e-07 | -1.66 | 9.69e-02 |
| DMPK[bd] | 19 | 45769717 | 45782552 | 16.52 | 2.79e-61 | -11.18 | 5.21e-29 |

*a* known GWAS risk gene of AD.
*b* gene within 1MB of known GWAS risk gene of AD.
*c* previously identified TWAS risk gene of AD.
*d* gene within 1MB of previously identified TWAS risk gene of AD.

**Table 3. Independent TWAS risk genes of Parkinson's disease identified by SR-TWAS.**

| | | | | SR-TWAS BRNSNG | | TIGAR BRNSNG | |
|---|---|---|---|---|---|---|---|
| **Gene** | **CHR** | **Start** | **End** | **Zscore** | **Pvalue** | **Zscore** | **Pvalue** |
| RAB29[ac] | 1 | 205767986 | 205775460 | -5.30 | 1.17e-07 | -4.17 | 3.06e-05 |
| SLC30A3 | 2 | 27254572 | 27275817 | -4.95 | 7.56e-07 | -0.74 | 4.62e-01 |
| IDUA[bd] | 4 | 986997 | 1004506 | 5.77 | 7.98e-09 | 4.05 | 5.09e-05 |
| CD38[ac] | 4 | 15778275 | 15853230 | 7.68 | 1.57e-14 | 2.10 | 3.59e-02 |
| SHROOM3[bd] | 4 | 76435100 | 76783253 | 5.17 | 2.32e-07 | 1.85 | 6.48e-02 |
| MMRN1[ac] | 4 | 89879532 | 89954629 | -7.91 | 2.55e-15 | -8.50 | 1.95e-17 |
| NDUFAF2[ac] | 5 | 60945129 | 61153037 | -5.45 | 4.91e-08 | -3.71 | 2.10e-04 |
| GPNMB[ac] | 7 | 23235967 | 23275108 | -5.52 | 3.31e-08 | -5.11 | 3.28e-07 |
| LA16c-385E7.1 | 16 | 1512979 | 1514675 | -4.81 | 1.50e-06 | -2.40 | 1.64e-02 |
| VKORC1[bc] | 16 | 31090842 | 31095980 | -5.56 | 2.75e-08 | -3.60 | 3.22e-04 |
| ZSWIM7[c] | 17 | 15976560 | 15993830 | -5.32 | 1.04e-07 | -4.13 | 3.59e-05 |
| LRRC37A4P[bd] | 17 | 45506741 | 45550335 | -9.41 | 4.98e-21 | -9.13 | 6.85e-20 |

*a* known GWAS risk gene of PD.
*b* gene within 1MB of known GWAS risk gene of PD.
*c* previously identified TWAS risk gene of PD.
*d* gene within 1MB of previously identified TWAS risk gene of PD.

## Figures

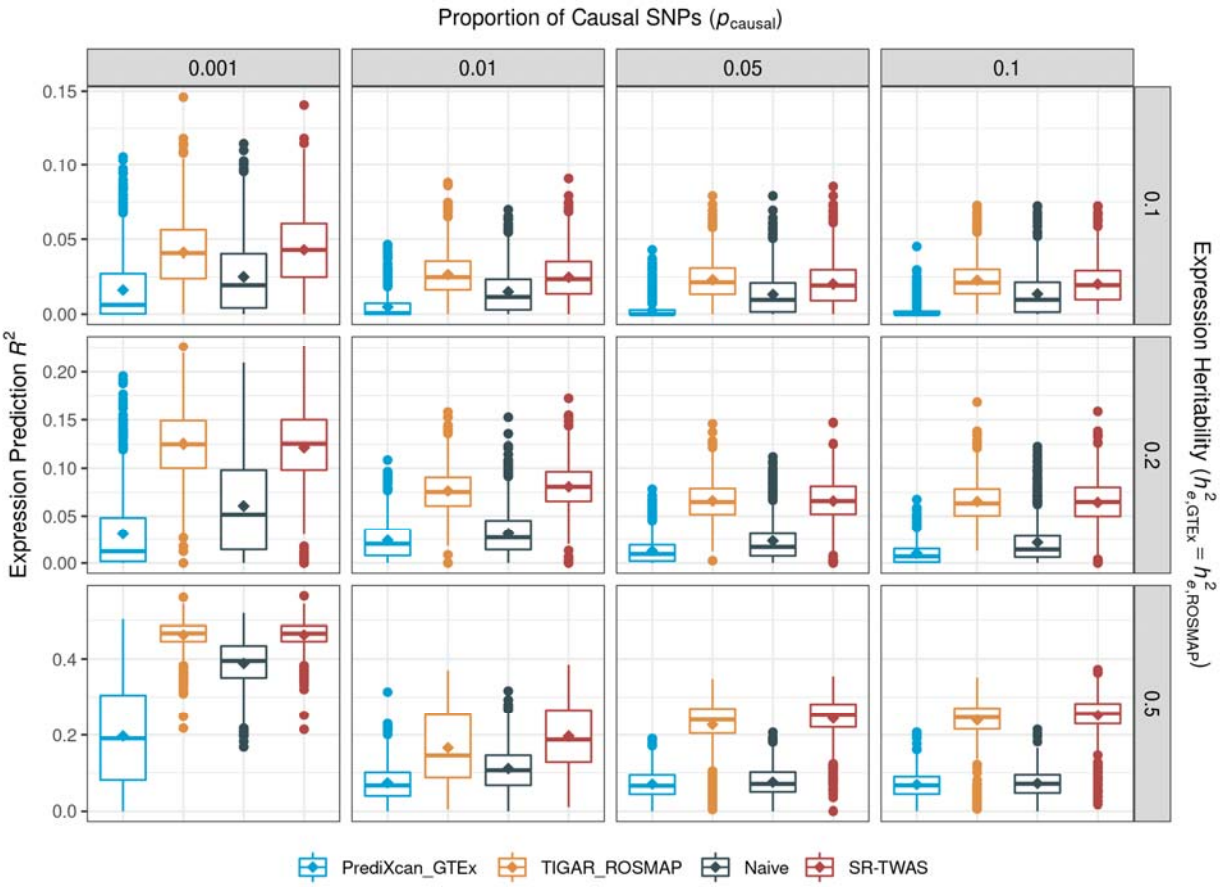


**Figure 1. Boxplots of gene expression prediction $R^2$ for simulations with varying proportion of true causal SNPs $p_{causal} = (0.001, 0.01, 0.05, 0.1)$ and true expression heritability $h^2_e = (0.1, 0.2, 0.5)$.** SR-TWAS performed comparably as the TIGAR_ROSMAP base model, but consistently better than the Naïve and PrediXcan_GTEx base models. This is because test samples are simulated under the same genetic architecture as the ROSMAP training cohort used by TIGAR_ROSMAP and the validation cohort used by SR-TWAS, which only have 50% overlapped true causal SNPs as the GTEx training cohort used by PrediXcan_GTEx.

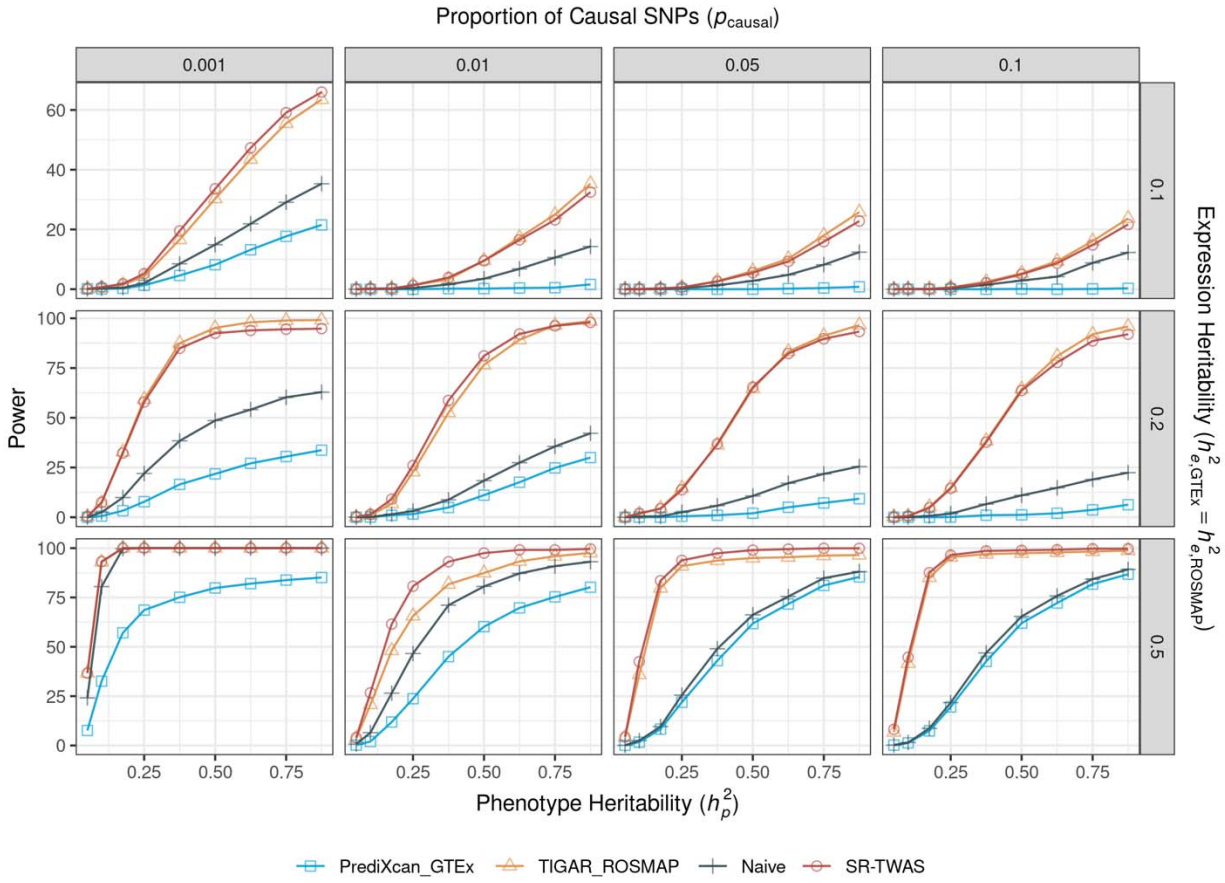


**Figure 2. Power comparison for simulations with varying proportion of true causal SNPs $p_{\text{causal}} = (0.001, 0.01, 0.05, 0.1)$, true expression heritability $h_e^2 = (0.1, 0.2, 0.5)$, and phenotype heritability $h_p^2 \in (0.05, 0.875)$.** SR-TWAS performed comparably as the TIGAR_ROSMAP base models trained using ROSMAP training cohort, but consistently better than the Naïve method and PrediXcan_GTEx base models trained using GTEx training cohort. This is because test samples are simulated under the same genetic architecture as the ROSMAP training cohort and the validation cohort used by SR-TWAS, which only have 50% overlapped true causal SNPs as the GTEx training cohort.

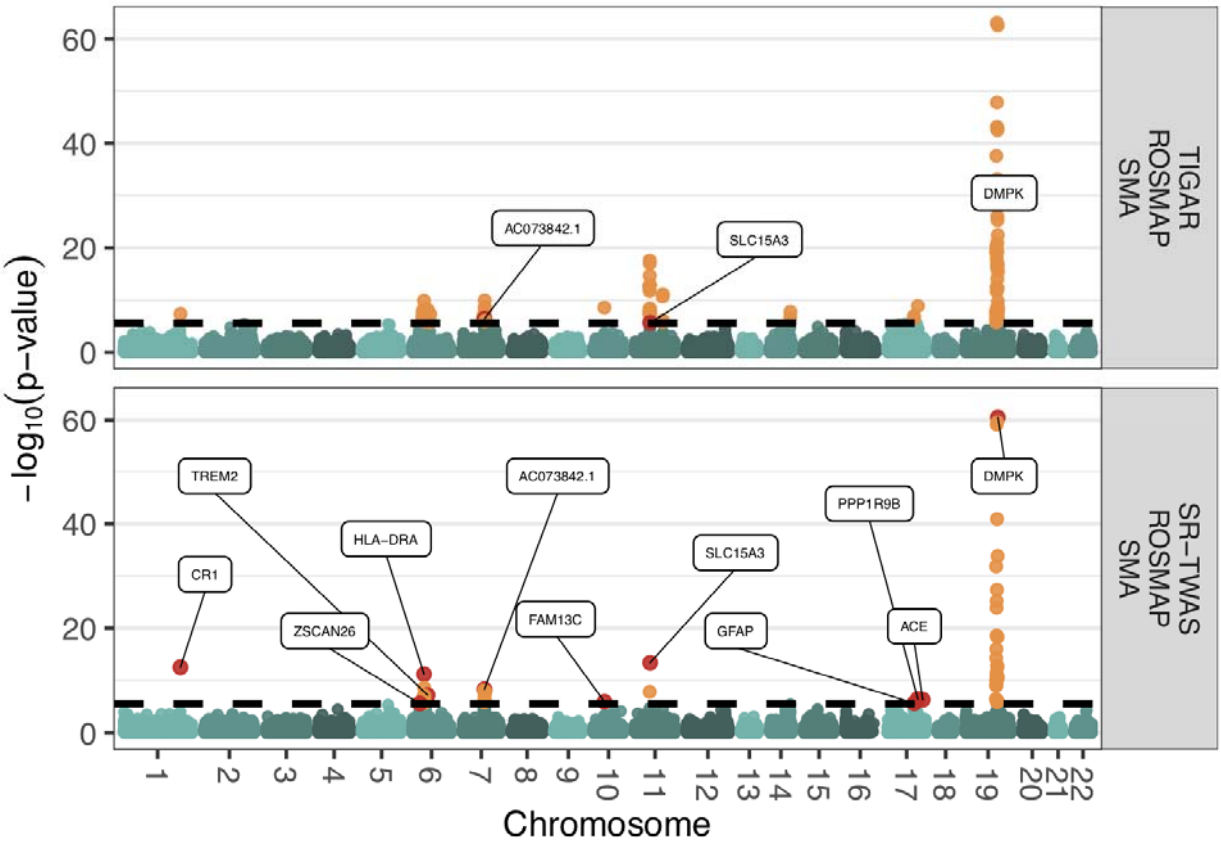


**Figure 3. Manhattan plots of TWAS results by SR-TWAS and TIGAR models trained for the target SMA tissue for studying AD dementia.** TIGAR models were trained using the validation data as used by SR-TWAS method. A total of 56 (11 independent) TWAS risk genes were identified by SR-TWAS, and 3 out of these 11 independent genes were also identified by TIGAR models. Significant genes are shown in orange and significant genes that are labeled and discussed in the text are shown in red.

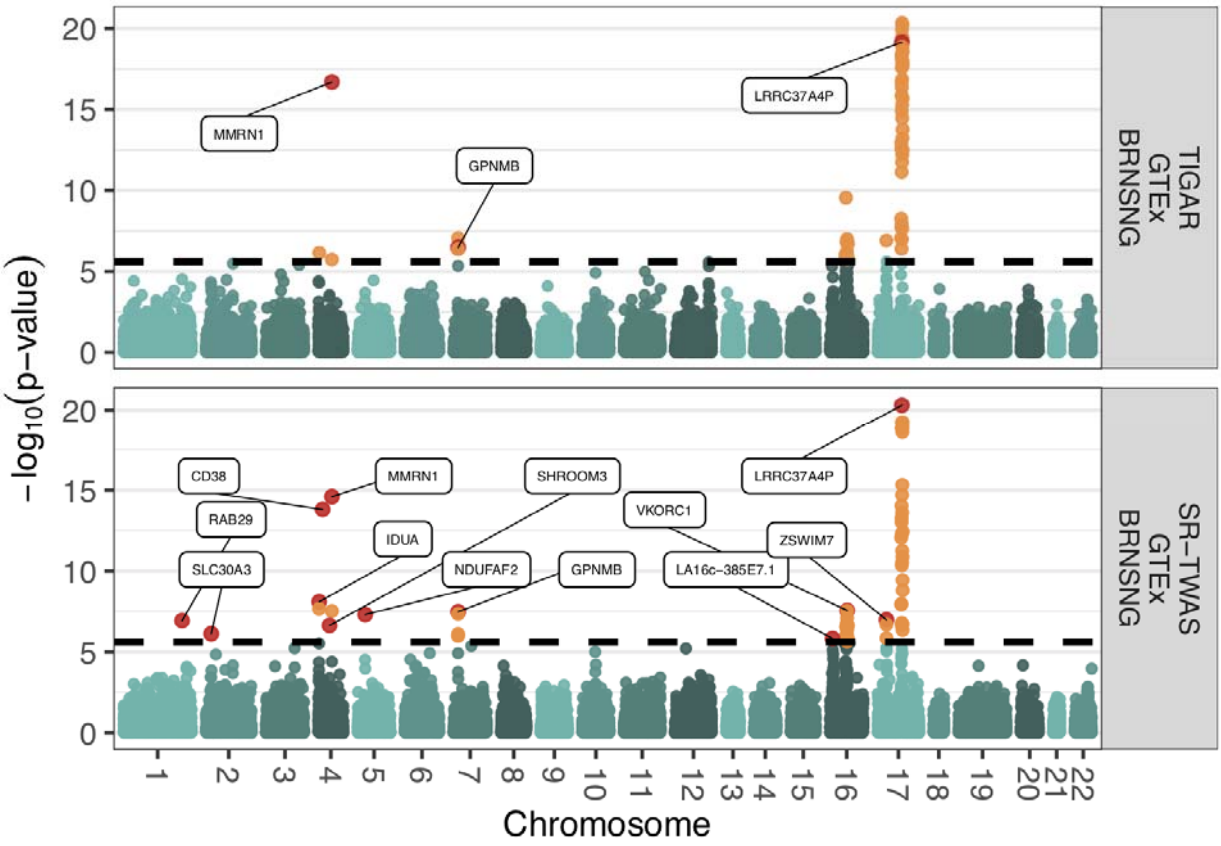


**Figure 4. Manhattan plots of TWAS results by SR-TWAS and TIGAR models trained for the target substantial nigral tissue for studying Parkinson's Disease.** TIGAR models were trained using the validation data as used by SR-TWAS method. A total of 62 (12 independent) significant TWAS risk genes were identified by SR-TWAS, and 3 of these 12 independent genes were also identified by TIGAR models. Significant genes are shown in orange and significant genes that are labeled and discussed in the text are shown in red.

# Methods

## SR-TWAS using Stacked Regression

Stacked regression is a machine learning method for forming optimal linear combinations of different predictors to improve prediction accuracy[14]. The theoretical background for combining predictors rather than selecting a single best predictor is well-established and has been developed since the 1970s[14,56,57]. The "stacking" method of combining predictors originated in a 1992 paper[13] by Wolpert, who described the concept as any scheme for feeding information from a set of cross-validated models to another before forming the final prediction in order to reduce prediction error[13]. The idea is further expanded with stacked regression, a specific framework for combining the initial predictors by weighted average with coefficient constraints to control for multicolinearity[14].

In standard two-stage TWAS, we need to first fit a gene expression imputation model, which is assumed as a multivariable linear regression model, with quantitative gene expression levels $E_g$ for the target gene and tissue type as the response variable, and genotype matrix $G$ of nearby/genome-wide SNPs as predictors,

$$\boldsymbol{E_g} = \boldsymbol{Gw} + \boldsymbol{\epsilon}, \qquad \epsilon_i \sim N(0,1).$$

This gene expression imputation model can be trained per gene per tissue type, using a transcriptomic reference panel which profiles both transcriptomic and genetic data of the same training cohort. SNPs with non-zero effect sizes $\boldsymbol{w}$ are referred to as a broad sense of eQTL. The eQTL effect sizes $\boldsymbol{w}$ will be estimated from each trained model by different regression methods and/or using different reference data of multiple tissue types.

Assume there are a total of $K$ base gene expression imputation models that are trained for the same target gene and tissue type, with $\widehat{\boldsymbol{w_k}}, k = 1, \ldots, K,$ as the trained eQTL effect sizes per base

519 model. Let $\mathbf{E}_{vg}$ denote the gene expression levels of the same target gene $g$ and tissue type in the
520 validation data, and $\mathbf{G}_v$ denote the genotype matrix of the same genetic predictors in the
521 validation data. Then the predicted Genetically Regulated gene eXpression (GReX) of the
522 validation samples are given by $\mathbf{G}_v\widehat{\mathbf{w}}_k$, by the $k$th base model. The stacked regression
523 method[13,14] will solve for a set of optimal model weights $\zeta_1, \ldots, \zeta_K$, by maximizing the regression
524 $R^2$ between the profiled gene expression $\mathbf{E}_{vg}$ and the weighted average GReX, $\sum_{k=1}^{K}\zeta_k\,\mathbf{G}_v\widehat{\mathbf{w}}_k$, of
525 $K$ base models, ie, minimizing the following loss function of $1-R^2$:

$$\text{minimize}_{(\zeta_k;\,k=1,\ldots,K)}\frac{\left\|\mathbf{E}_{vg}-\sum_{k=1}^{K}\zeta_k\,\mathbf{G}_v\widehat{\mathbf{w}}_k\right\|^2}{\left\|\mathbf{E}_{vg}-E_{vg}\right\|^2},\ \text{s.t.}\ \sum_{k=1}^{K}\zeta_k=1,\ \zeta_k\in[0,1].$$

526 As a result, we will obtain a set of model weights $\zeta_k$ for $k=1,\ldots,K$ base models, and a set of
527 eQTL effect sizes $\widetilde{\mathbf{w}}$ given by the weighted average of the eQTL effect sizes of $K$ base models,
528 $\widetilde{\mathbf{w}}=\sum_{k=1}^{K}\zeta_k\,\widehat{\mathbf{w}}_k$ (Stage I). Then the final predicted GReX for test genotype data $\mathbf{G}_t$ is given by
529 $\widehat{\text{GReX}}_g=\mathbf{G}_t\widetilde{\mathbf{w}}$, and $\widetilde{\mathbf{w}}$ will be taken as variant weights in the gene-based association tests by
530 SR-TWAS in Stage II.
531 Genes with 5-fold CV $R^2>0.5\%$ in the validation dataset by SR-TWAS are considered as
532 having a valid imputation model and will be tested in Stage II. That is, the validation dataset will
533 be randomly split into 5 folds. For each fold of data, SR-TWAS model will be trained using the
534 other 4-fold data and then use to calculate prediction $R^2$ with the current fold. The average
535 prediction $R^2$ across all 5 folds of data is considered as the 5-fold CV $R^2$. Here, we use a more
536 liberal threshold (0.005) than the threshold 0.01 used by previous studies [15,58,59] to allow more
537 genes to be tested in follow-up TWAS. Because the follow-up gene-based association Z-score
538 test statistic is essentially a weighted average of single variant GWAS Z-score statistics with

variant weights provided by the eQTL effect sizes[8], poorly estimated eQTL weights would only reduce power but will not increase false positive rate under null hypothesis.

**Naïve Method**

In this paper, we compared SR-TWAS to a Naïve approach which just takes the average of base models as the trained gene expression imputation model, that is, takes $\zeta_k = \frac{1}{K}, k = 1, \ldots, K$. Using a validation dataset, we can still evaluate the validation $R^2$ which can be used to select valid genes with validation $R^2 > 0.5\%$.

**SR-TWAS Tool Framework**

SR-TWAS tool was designed to be compatible with the TIGAR-V2 tool framework[8]; it accepts models trained by TIGAR-V2 as input, imports utility functions from TIGAR-V2, and outputs model files which can be used as input for TIGAR-V2 GReX prediction and summary-level TWAS. Much of the structure of the SR-TWAS code was derived from existing TIGAR-V2 scripts and it shares dependencies on TABIX[60] and the Python libraries of numpy[61,62], pandas[61], scipy[63], statsmodels[64], and scikit-learn[65,66].

The SR-TWAS script utilizes scikit-learn's consistent, extensible interfaces for defining estimators and predictors and for initializing objects[66]. The script trains a stacked regression model using a modified version of scikit-learn's StackingRegressor class, which trains a final estimator from cross-validated predictions from base estimators fitted on the full design matrix. The script defines two custom classes to be used as input for the stacking regressor object: a base estimator class (WeightEstimator) which converts trained GReX prediction models into scikit-learn-compatible estimator objects and a final estimator class (ZetasEstimator) which obtains the

values of $\zeta_1, \ldots, \zeta_K$ that minimize the loss function under the constraints $\zeta_k \geq 0$ and $\sum_{k=1}^{K} \zeta_k = 1$[14].

During the stacked regression, SNP minor allele frequencies and effect sizes for the specified target are first read from each of the $K$ user-specified weight files. The SNPs are then matched to SNPs in the validation genotype data and filtered to exclude effect sizes of SNPs for which the difference between the MAF of the genotype data and the MAF from the corresponding weight file exceeds a user-specified MAF difference threshold. The effect sizes from each weight file are used to initialize $K$ separate instances of the WeightEstimator class. These $K$ WeightEstimator objects are used as base estimators and fit on genotype and expression data from the validation data.

Only SR-TWAS models trained from $K = 2, 4, 6$ base models are presented in this paper. The code was designed to accept any $K \geq 2$, and while the stacked regression script has been primarily tested using $K = 2, 4, 6$ base models, preliminary testing with dummy weight files confirms it can train stacked regression models from $K > 6$ base models.

**ROS/MAP Reference Panel**

The Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP) are two ongoing longitudinal, epidemiologic clinical-pathologic cohort studies of aging and Alzheimer's disease collectively referred to as ROS/MAP[10]. ROS enrolls Catholic nuns, priests, and brothers from religious groups across the United States, primarily from communal living settings[10]. While the similar adult lifestyle of participants allows for more control of potential confounders such as education and socioeconomic status, it simultaneously limits the ability to study such variables[10].

584 MAP was designed to complement and extend studies like ROS by including subjects from a
585 wider range of life experiences, socioeconomic status, and educational attainment and recruits
586 participants primarily from retirement communities in the Chicago area, but also subsidized
587 housing, retirement homes, and through organizations serving minorities and low-income
588 elderly[10]. All participants in both studies are without known dementia and agree to annual
589 clinical evaluations and brain donation upon death[10]. Similarity in study design and data
590 collection procedures allows the ROS and MAP datasets to be merged for use in joint
591 analyses[10,67].
592 Quality-controlled ROS/MAP WGS data for European subjects[67] were used for both the real data
593 application and simulation studies. Transcriptomic data of ROS/MAP samples of brain PFC were
594 profiled by RNA-sequencing (RNA-seq). Gene expression data of Transcripts Per Million (TPM)
595 per sample were provided by Rush Alzheimer's Disease Center. Genes with $> 0.1$ TPM in $\geq 10$
596 samples were considered. Raw gene expression data (TPM) were then log2 transformed and
597 adjusted for age at death, sex, postmortem interval, study (ROS or MAP), batch effects, RNA
598 integrity number scores, cell type proportions (with respect to oligodendrocytes, astrocytes,
599 microglia, neurons), top five genotype principal components, and top probabilistic estimation of
600 expression residuals (PEER) factors[68] by linear regression models. SNPs with minor allele
601 frequency (MAF) > 1%, Hardy-Weinberg p-value $> 10^{-5}$ were analyzed. For each gene, cis-
602 SNPs within the 1MB of the flanking 5' and 3' ends were used in the imputation models as
603 predictors.
604
605 **GTEx V8 Reference Panel**

606 The Genotype-Tissue Expression (GTEx) project V8 profiles both whole genome sequencing
607 (WGS) genotype data and RNA-seq transcriptomic data for Brain PFC tissue type of n=157
608 donors[9]. Gene expression data of Transcripts Per Million (TPM) per sample per tissue were
609 downloaded from the GTEx portal. Genes with $> 0.1$ TPM in $\geq 10$ samples were considered.
610 Raw gene expression data (TPM) were then log2 transformed and adjusted for age, body mass
611 index (BMI), top five genotype principal components, and top probabilistic estimation of
612 expression residuals (PEER) factors[68]. SNPs with minor allele frequency (MAF) > 1%, Hardy-
613 Weinberg p-value $> 10^{-5}$ were analyzed. For each gene, cis- SNPs within the 1MB of the
614 flanking 5' and 3' ends were used in the imputation models as predictors.
615
616 **Simulation Study Design**
617 We conducted in depth simulation studies under various scenarios to assess the performance of
618 SR-TWAS, Naïve method, and base models by PrediXcan and TIGAR. We used the real
619 genotype data of gene *ABCA7* from ROS/MAP and GTEx V8 to simulate gene expression and
620 phenotypes. We considered three different settings: (i) Samples from ROS/MAP and GTEx
621 cohorts have the same set of true causal SNPs (i.e., the same genetic architecture). The
622 expression heritability was the same for both ROS/MAP and GTEx V8 cohorts. (ii) Samples
623 from ROS/MAP and GTEx cohorts have the same set of true causal SNPs (i.e., the same genetic
624 architecture). The expression heritability for GTEx V8 cohort is only half of the one for
625 ROS/MAP. (iii) Samples from the ROS/MAP cohort were simulated with the same causal SNPs
626 (i.e., eQTL), while samples from the GTEx V8 cohort were simulated with true causal SNPs that
627 were 50% overlapped with the ones for ROS/MAP. The expression heritability was the same for
628 both ROS/MAP and GTEx V8 cohorts.

Under each setting, we considered multiple scenarios with varying proportions of causal SNPs ($p_{causal} = (0.001, 0.01, 0.05, 0.1)$) and gene expression heritability (i.e., the proportion of gene expression variation due to genetics, $h_e^2 = (0.1, 0.2, 0.5)$). We randomly selected n=465 training samples with WGS genotype data from ROS/MAP and GTEx V8, respectively. We randomly selected n=400 and n=800 samples with WGS genotype data from ROS/MAP as our validation and test cohorts, respectively. We considered a series of $h_p^2$ values, the proportion of phenotype variance due to simulated gene expression, in the range of $(0.05, 0.875)$.

For each scenario, gene expression $\mathbf{E}_i$ for the $ith$ simulation iteration is generated using the following formula

$$\mathbf{E}_i = \gamma_i \mathbf{G}^* \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \;\; \gamma_i = \sqrt{\frac{h_e^2}{\mathrm{Var}(\mathbf{G}^* \boldsymbol{\beta}_{\bullet i})}}, \;\; \boldsymbol{\varepsilon}_i \sim \mathrm{N}\left(0, \sqrt{1 - h_e^2}\right),$$

where $\mathbf{G}^*$ denotes the genotype matrix of $N_{\text{causal}}$ randomly chosen true causal SNPs for all samples, effect size vector $\boldsymbol{\beta}_\mathbf{i}$ was generated from $N(0, I)$, and $\gamma_i$ is a scale factor chosen to ensure the targeted $h_e^2$ value. The phenotype vector $\mathbf{Y}_i$ for the $i$th simulation iteration was generated using the following formula

$$\mathbf{Y}_i = \varphi_i \mathbf{E}_i + \boldsymbol{\varepsilon}_i, \;\; \varphi_i = \sqrt{\frac{h_p^2}{\mathrm{Var}(\mathbf{E}_i)}}, \;\; \boldsymbol{\varepsilon}_i \sim \mathrm{N}\left(0, \sqrt{1 - h_p^2}\right)$$

where $\mathbf{E}_i$ is the simulated gene expression, and $\varphi_i$ is a scale factor to ensure the targeted $h_p^2$ value.

Two base models per gene were trained by PrediXcan with the GTEx training samples (n=465), and by TIGAR with the ROS/MAP training samples (n=465). SR-TWAS and Naïve models were then obtained by using these trained base models. Validation data ($n = 400$) were used to train SR-TWAS models and filter out gene expression imputation models with 5-fold cross-

validation $R^2 < 0.5\%$ in the validation cohort for both SR-TWAS and Naïve models. Test data (n=800) were used for assessing GReX prediction performance and TWAS power. Each causal simulation scenario was repeated for 1,000 times. We compared the performance by SR-TWAS, Naïve method, and these two base models with respect to prediction imputation $R^2$ in the test data and the power of TWASs.

The predicted $\widehat{\mathrm{GReX}}_i$ by each trained gene expression imputation model was used to calculate expression prediction $R^2$, which is equivalent to the regression $R^2$ between profiled and predicted gene expression, given by

$$R^2_{E_i} = \mathrm{Cor}\left(\mathbf{E}_i, \widehat{\mathrm{GReX}}_i\right)^2.$$

The power will be given by the proportion of simulation iterations that have TWAS p-value $< 2.5 \times 10^6$ out of a total of 1,000 simulation iterations.

## Data availability

All ROS/MAP data analyzed in this study are de-identified and available to any qualified investigator with application through the Rush Alzheimer's Disease Center Research Resource Sharing Hub, https://www.radc.rush.edu, which has descriptions of the studies and available data. GTEx V8 data are available from dbGaP with accession phs000424.v8.p2. TIGAR_GTEx base models trained from GTEx V8 are available at Synapse https://www.synapse.org/TIGAR_V2_Resource_GTExV8. PrediXcan_GTEx base models trained from GTEx V8 are available from https://predictdb.org/. GWAS summary data of AD are available from https://ctg.cncr.nl/software/summary_statistics, and GWAS summary data of PD are available from https://bit.ly/2ofzGrk. TIGAR_ROSMAP and PrediXcan_ROSMAP base models trained from ROS/MAP, trained SR-TWAS and Naïve models of brain PFC tissue type

in this study, and all TWAS summary statistics will be made available at SYNAPSE when the manuscript is accepted for publication.

## Code availability

The SR-TWAS tool, including the Naïve method as an option, is publicly available on GitHub, https://github.com/yanglab-emory/SR-TWAS.

677 **References**

678 1. Feng, H. *et al.* Transcriptome‐wide association study of breast cancer risk by
679 estrogen‐receptor status. *Genet. Epidemiol.* **44**, 442–468 (2020).
680 2. Kar, S. *et al.* Pleiotropy-guided transcriptome imputation from normal and tumor tissues
681 identifies new candidate susceptibility genes for breast and ovarian cancer. *bioRxiv* (2020)
682 doi:10.1101/2020.04.23.043653.
683 3. Strunz, T., Lauwen, S., Kiel, C., Hollander, A. den & Weber, B. H. F. A transcriptome-
684 wide association study based on 27 tissues identifies 106 genes potentially relevant for
685 disease pathology in age-related macular degeneration. *Scientific Reports* **10**, 1584 (2020).
686 4. Wu, C. *et al.* Transcriptome-wide association study identifies susceptibility genes for
687 rheumatoid arthritis. *Arthritis Res Ther* **23**, (2021).
688 5. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association
689 studies. *Nat Genet* **51**, 592–599 (2019).
690 6. Nagpal, S. *et al.* TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation
691 Enhances Gene Mapping of Complex Traits. *The American Journal of Human Genetics*
692 **105**, 258–266 (2019).
693 7. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene
694 expression variation inferred from GWAS summary statistics. *Nat Commun* **9**, 1–20 (2018).
695 8. Parrish, R. L., Gibson, G. C., Epstein, M. P. & Yang, J. TIGAR-V2: Efficient TWAS tool
696 with nonparametric Bayesian eQTL weights of 49 tissue types from GTEx V8. *Human*
697 *Genetics and Genomics Advances* **3**, 100068 (2022).
698 9. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human
699 tissues. *Science* **369**, 1318–1330 (2020).

10. Bennett, D. A. *et al.* Religious Orders Study and Rush Memory and Aging Project. *J Alzheimers Dis* **64**, S161–S189 (2018).

11. Hu, Y. *et al.* A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat Genet* **51**, 568–576 (2019).

12. Liu, A. E. & Kang, H. M. Meta-imputation of transcriptome from genotypes across multiple datasets by leveraging publicly available summary-level data. *PLOS Genetics* **18**, e1009571 (2022).

13. Wolpert, D. H. Stacked generalization. *Neural Networks* **5**, 241–259 (1992).

14. Breiman, L. Stacked regressions. *Mach Learn* **24**, 49–64 (1996).

15. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**, 245–252 (2016).

16. Wightman, D. P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat Genet* **53**, 1276–1282 (2021).

17. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *The American Journal of Human Genetics* **100**, 473–487 (2017).

18. Marioni, R. E. *et al.* GWAS on family history of Alzheimer's disease. *Transl Psychiatry* **8**, 99 (2018).

19. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet* **51**, 404–413 (2019).

20. Nazarian, A., Yashin, A. I. & Kulminski, A. M. Genome-wide analysis of genetic predisposition to Alzheimer's disease and related sex disparities. *Alz Res Therapy* **11**, 1–21 (2019).

21. Gockley, J. *et al.* Multi-tissue neocortical transcriptome-wide association study implicates 8 genes across 6 genomic loci in Alzheimer's disease. *Genome Medicine* **13**, 76 (2021).
22. Jing, Q. *et al.* A Comprehensive Analysis Identified Hub Genes and Associated Drugs in Alzheimer's Disease. *BioMed Research International* **2021**, e8893553 (2021).
23. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet* **51**, 414–430 (2019).
24. Schwartzentruber, J. *et al.* Genome-wide meta-analysis, fine-mapping, and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat Genet* **53**, 392–402 (2021).
25. Lambert, J.-C. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet* **41**, 1094–1099 (2009).
26. Jonsson, T. *et al.* Variant of TREM2 Associated with the Risk of Alzheimer's Disease. *N Engl J Med* **368**, 107–116 (2013).
27. Tang, H. & Harte, M. Investigating Markers of the NLRP3 Inflammasome Pathway in Alzheimer's Disease: A Human Post-Mortem Study. *Genes (Basel)* **12**, 1753 (2021).
28. Shigemizu, D. *et al.* Ethnic and trans-ethnic genome-wide association studies identify new loci influencing Japanese Alzheimer's disease risk. *Transl Psychiatry* **11**, 1–10 (2021).
29. Deming, Y. *et al.* The MS4A gene cluster is a key modulator of soluble TREM2 and Alzheimer's disease risk. *Science Translational Medicine* **11**, eaau2291 (2019).
30. Jimenez□Marin, A. *et al.* Transcriptional signatures of synaptic vesicle genes define myotonic dystrophy type I neurodegeneration. *Neuropathol Appl Neurobiol* **47**, 1092–1108 (2021).

31. Burdelski, C. *et al.* Family with sequence similarity 13C (FAM13C) overexpression is an independent prognostic marker in prostate cancer. *Oncotarget* **8**, 31494–31508 (2017).

32. Lin, H.-C. *et al.* Alzheimer's disease is associated with prostate cancer: a population-based study. *Oncotarget* **9**, 7616–7622 (2018).

33. Pereira, J. B. *et al.* Plasma GFAP is an early marker of amyloid-β but not tau pathology in Alzheimer's disease. *Brain* **144**, 3505–3516 (2021).

34. Akram, A. *et al.* Stereologic estimates of total spinophilin-immunoreactive spine number in area 9 and the CA1 field: Relationship with the progression of Alzheimer's disease. *Neurobiology of Aging* **29**, 1296–1307 (2008).

35. Mi, Z. *et al.* Loss of precuneus dendritic spines immunopositive for spinophilin is related to cognitive impairment in early Alzheimer's disease. *Neurobiology of Aging* **55**, 159–166 (2017).

36. Wu, M. *et al.* Hippocampal overexpression of TREM2 ameliorates high fat diet induced cognitive impairment and modulates phenotypic polarization of the microglia. *Genes & Diseases* **9**, 401–414 (2022).

37. Song, F. *et al.* Regulation and biological role of the peptide/histidine transporter SLC15A3 in Toll-like receptor-mediated inflammatory responses in macrophage. *Cell Death Dis* **9**, 1–15 (2018).

38. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet Neurology* **18**, 1091–1102 (2019).

39. Yao, S. *et al.* A transcriptome-wide association study identifies susceptibility genes for Parkinson's disease. *npj Parkinsons Dis.* **7**, 1–8 (2021).

40. Kia, D. A. *et al.* Identification of Candidate Parkinson Disease Genes by Integrating Genome-Wide Association Study, Expression, and Epigenetic Data Sets. *JAMA Neurology* **78**, 464–472 (2021).

41. Robak, L. A. *et al.* Excessive burden of lysosomal storage disorder gene variants in Parkinson's disease. *Brain* **140**, 3191–3203 (2017).

42. Subrahmanian, N. & LaVoie, M. J. Is there a special relationship between complex I activity and nigral neuronal loss in Parkinson's disease? A critical reappraisal. *Brain Research* **1767**, 147434 (2021).

43. Storm, C. S. *et al.* Finding genetically-supported drug targets for Parkinson's disease using Mendelian randomization of the druggable genome. *Nat Commun* **12**, 7342 (2021).

44. Prokop, J. W. *et al.* Characterization of Coding/Noncoding Variants for SHROOM3 in Patients with CKD. *J Am Soc Nephrol* **29**, 1525–1535 (2018).

45. Meléndez-Flores, J. D. & Estrada-Bellmann, I. Linking chronic kidney disease and Parkinson's disease: a literature review. *Metab Brain Dis* **36**, 1–12 (2021).

46. Sikora, J. & Ouagazzal, A.-M. Synaptic Zinc: An Emerging Player in Parkinson's Disease. *Int J Mol Sci* **22**, 4724 (2021).

47. Guerreiro, S., Privat, A.-L., Bressac, L. & Toulorge, D. CD38 in Neurodegeneration and Neuroinflammation. *Cells* **9**, 471 (2020).

48. Moloney, E. B., Moskites, A., Ferrari, E. J., Isacson, O. & Hallett, P. J. The glycoprotein GPNMB is selectively elevated in the substantia nigra of Parkinson's disease patients and increases after lysosomal stress. *Neurobiology of Disease* **120**, 1–11 (2018).

49. Purlyte, E. *et al.* Rab29 activation of the Parkinson's disease-associated LRRK2 kinase. *The EMBO Journal* **37**, 1–18 (2018).

50. Sagi, O. & Rokach, L. Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery* **8**, e1249 (2018).

51. Mogil, L. S. *et al.* Genetic architecture of gene expression traits across diverse populations. *PLOS Genetics* **14**, e1007586 (2018).

52. Luningham, J. M. *et al.* Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. *The American Journal of Human Genetics* **107**, 714–726 (2020).

53. Yang, C. *et al.* CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. *Bioinformatics* **35**, 1644–1652 (2019).

54. Yuan, Z. *et al.* Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nature Communications* **11**, 3861 (2020).

55. Tang, S. *et al.* Novel Variance-Component TWAS method for studying complex human diseases with applications to Alzheimer's dementia. *PLOS Genetics* **17**, e1009482 (2021).

56. Rao, J. N. K. & Subrahmaniam, K. Combining Independent Estimators and Estimation in Linear Regression with Unequal Variances. *Biometrics* **27**, 971–990 (1971).

57. Efron, B. & Morris, C. Combining Possibly Related Estimation Problems. *Journal of the Royal Statistical Society* **35**, 379–421 (1973).

58. Wu, L. *et al.* A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat Genet* **50**, 968–978 (2018).

59. Bhattacharya, A., Li, Y. & Love, M. I. MOSTWAS: Multi-Omic Strategies for Transcriptome-Wide Association Studies. *PLoS Genet* **17**, e1009398 (2021).

60. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719 (2011).

61. McKinney, W. Data Structures for Statistical Computing in Python. in *Proceedings of the 9th Python in Science Conference (SciPy 2010)* 56–61 (2010). doi:10.25080/Majora-92bf1922-00a.

62. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

63. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).

64. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. in *Proceedings of the 9th Python in Science Conference (SciPy 2010)* 92–96 (2010). doi:10.25080/Majora-92bf1922-011.

65. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

66. Buitinck, L. *et al.* API design for machine learning software: experiences from the scikit-learn project. in *Proceedings of the European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECMPKDD'13)* 108--122 (2013).

67. De Jager, P. L. *et al.* A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci Data* **5**, (2018).

68. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500–507 (2012).